January 1989 LIDS-P-1845

Simulated Annealing Type Algorithms for Multivariate Optimization[1]

by

Saul B. Gelfand[2] and Sanjoy K. Mitter[3]

Abstract

We study the convergence of a class of discrete-time continuous-state stimulated annealing type algorithms for multivariate optimization. The general algorithm that we consider is of the form $X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k$. Here $U(\bullet)$ is a smooth function on a compact subset of $\mathbb{R}^r$, $\{\xi_k\}$ is a sequence of $\mathbb{R}^r$ - valued random variables, $\{W_k\}$ is a sequence of independent standard r-dimensional Gaussian random variables, and $\{a_k\}$, $\{b_k\}$ are sequences of positive numbers which tend to zero. These algorithms arise by adding slowly decreasing white Gaussian noise to gradient descent, random search, and stochastic approximation algorithms. We show that under suitable conditions on $U(\bullet)$, $\{\xi_k\}$, $\{a_k\}$ and $\{b_k\}$ that X_k converges in probability to the set of global minima of $U(\bullet)$.

[1]Research reported here has been supported by the Air Force Office of Scientific Research under grant AFOSR-85-0227B.

[2]Computer Vision and Image Processing Laboratory, School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

[3]Laboratory for Information and Decision Systems, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **JAN 1989**	2. REPORT TYPE	3. DATES COVERED **00-01-1989 to 00-01-1989**

4. TITLE AND SUBTITLE **Simulated Annealing Type Algorithms for Multivariate Optimization**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Laboratory for Information and Decision Systems,77 Massachusetts Avenue,Cambridge,MA,02139-4307**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**22**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

1. INTRODUCTION

It is desired to select a parameter value x^* which minimizes a smooth function $U(x)$ over $x \in D$, where D is compact subset of $\mathbb{R}^r$. The stochastic descent algorithm

$$Z_{k+1} = Z_k - a_k(\nabla U(Z_k) + \xi_k) \; , \tag{1.1}$$

is often used where $\{\xi_k\}$ is a sequence of $\mathbb{R}^r$ - valued random variables and $\{a_k\}$ is a sequence of positive numbers with $a_k \to 0$ and $\sum a_k = \infty$. An algorithm of this type might arise in several ways. The sequence $\{Z_k\}$ could correspond to a stochastic approximation [1], where the sequence $\{\xi_k\}$ arises from noisy measurements of $\nabla U(\cdot)$ or $U(\cdot)$. The sequence $\{Z_k\}$ could also correspond to a random search [2], where the sequence $\{\xi_k\}$ arises from randomly selected search directions. Now since D is compact it is necessary to insure the trajectories of $\{Z_k\}$ are bounded; this may be done either by projecting Z_k back into D if it ever leaves D, or by fixing the dynamics in (1.1) so that Z_k never leaves D or only leaves D finitely many times w.p.1. Let S be the set of local minima of $U(\cdot)$ and S^* the set of global minima of $U(\cdot)$. Under suitable conditions on $U(\cdot)$, $\{\xi_k\}$ and $\{a_k\}$, and assuming that $\{Z_k\}$ is bounded, it is well-known that $Z_k \to S$ as $k \to \infty$ w.p.1. In particular, if $U(\cdot)$ is well-behaved, $a_k = A/k$ for k large, and $\{\xi_k\}$ are independent random variables such that $E\{|\xi_k|^2\} \leq c\, a_k^\alpha$ and $|E\{\xi_k\}| \leq c\, a_k^\beta$ where $\alpha > -1$, $\beta > 0$, and c is a positive constant, then $Z_k \to S$ as $k \to \infty$ w.p.1. However, if $U(\cdot)$ has strictly local minima, then in general $Z_k \not\to S^*$ as $k \to \infty$ w.p.1.

The analysis of the convergence w.p.1 of $\{Z_k\}$ is usually based on the convergence of an *associated ordinary differential equation* (ODE)

$$\dot{z}(t) = -\nabla U(z(t)).$$

This approach was pioneered by Ljung [3] and further developed by Kushner and Clark [4], Metivier and Priouret [5], and others. Kushner and Clark also analyzed the convergence in probability of $\{Z_k\}$ by this method. However, although their theory yields much useful information about the asymptotic behavior of $\{Z_k\}$ under very weak assumptions, it fails to obtain $Z_k \to S^*$ as $k \to \infty$ in probability unless S is a singleton; see [4, p. 125].

Consider a modified stochastic descent algorithm

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k \qquad (1.2)$$

where $\{W_k\}$ is a sequence of independent Gaussian random variables with zero-mean and identity covariance matrix, and $\{b_k\}$ is a sequence of positive numbers with $b_k \to 0$. The $b_k W_k$ term is added in artificially by Monte Carlo simulation so that $\{X_k\}$ can avoid getting trapped in a strictly local minimum of $U(\cdot)$. In general $X_k \not\to S^*$ as $k \to \infty$ w.p.1 (for the same reasons that $Z_k \not\to S^*$ as $k \to \infty$ w.p.1). However, under suitable conditions on $U(\cdot)$, $\{\xi_k\}$, $\{a_k\}$ and $\{b_k\}$, and assuming that $\{X_k\}$ is bounded, we shall show that $X_k \to S^*$ as $k \to \infty$ in probability. In particular, if $U(\cdot)$ is well-behaved, $a_k = A/k$ and $b_k^2 = B/k \log\log k$ for k large where $B/A > C_0$ (a positive constant which depends only on $U(\cdot)$), and $\{\xi_k\}$ are independent random variables such that $E\{|\xi_k|^2\} \leq c\,a_k^\alpha$ and $|E\{\xi_k\}| \leq c\,a_k^\beta$ where $\alpha > -1$, $\beta > 0$, and c is a positive constant, then $X_k \to S^*$ as $k \to \infty$ in probability.

Our analysis of the convergence in probability of $\{X_k\}$ is based on the convergence of what we will call the *associated stochastic differential equation* (SDE)

$$dx(t) = -\nabla U(x(t))dt + c(t)dw(t) \qquad (1.3)$$

where $w(\cdot)$ is a standard r-dimensional Wiener process and $c(\cdot)$ is a positive function with $c(t) \to 0$ as $t \to \infty$ (take $t_k = \sum_{n=0}^{k-1} a_n$ and $b_k = \sqrt{a_k}\, c(t_k)$ to see the relationship between (1.2) and (1.3)). The simulation of the Markov diffusion $x(\cdot)$ for the purpose of global optimization has been called continuous simulated annealing. In this context, $U(x)$ is called the energy of state x and $T(t) = c^2(t)/2$ is called the temperature at time t. This method was first suggested by Grenender [6] and Geman and Hwang [7] for image processing applications with continuous grey levels. We remark that the discrete simulated annealing algorithm for combinatorial optimization based on simulating a Metropolis-type Markov chain [8], and the continuous simulated annealing algorithm for multivariate optimization based on simulating the Langevin-type Markov diffusion discussed above both have a (Gibbs) invariant distribution $\propto \exp(-U(x)/T)$ when the temperature is fixed at T. The invariant distributions concentrate on the global minima of $U(\cdot)$ as $T \to 0$. The discrete and continuous algorithms are further related in that a certain parametric family of continuous state Metropolis-type Markov chains interpolated into continuous time Markov processes converge to a Langevin-type Markov diffusion [9]. Now the asymptotic behavior of $x(\cdot)$ has been studied intensively by a number of researchers [7], [10]-[12]. Our work is based on the analysis of $x(\cdot)$ developed by Chiang, Hwang and Sheu [11] who prove the following result: if $U(\cdot)$ is well-behaved and $c^2(t) = C/\log t$ for t large where $C > C_0$ (a positive constant which depends only on $U(\cdot)$ and the same C_0 as above) then $x(t) \to S^*$ as $t \to \infty$ in probability.

The actual implementation of (1.3) on a digital computer requires some type of discretization or numerical integration, such as (1.2). Aluffi-Pentini, Parisi, and Zirilli [13] describe some numerical experiments performed with (1.2) for a variety of test problems. Kushner [12] was the first to analyze (1.2) but for the case of $a_k = b_k = A/\log k$, k large. Although Kushner obtains a detailed asymptotic description of $\{X_k\}$ for this case, in general $X_k \not\to S^*$ as $k \to \infty$ in probability unless $\xi_k = 0$. The reason for this is intuitively clear: even if $\{\xi_k\}$ is bounded, $a_k \xi_k$ and $a_k W_k$ can be of the same order and hence can interfere with each other. On the other hand by considering (1.2) for the case of $a_k = A/k$, $b_k^2 = B/k \log\log k$, k large, we get $X_k \to S^*$ as $k \to \infty$ in probability for $\{\xi_k\}$ with *unbounded* variance, in particular for $E\{|\xi_k|^2\} = O(k^\gamma)$ and $\gamma < 1$. Our method of analysis is different from Kushner's in that we obtain the asymptotic behavior of $\{X_k\}$ from $x(\cdot)$.

2. MAIN RESULTS AND DISCUSSION

We will use the following notation. If $F \subset \mathbb{R}^r$ then $\overset{\circ}{F}$ is the interior of F and ∂F is the boundary of F. $1_G(\cdot)$ is the indicator function for the set G. $|\cdot|$ and $\langle \cdot,\cdot \rangle$ are the Euclidean norm and inner product, respectively.

Our analysis, like Kushner's [12], requires that we bound the trajectories of $\{X_k\}$. We proceed as follows. Take D to be a closed ball in $\mathbb{R}^r$ centered at the origin. Let D_1 be another closed ball in $\mathbb{R}^r$ centered at the origin with $D_1 \subset D$ (strictly). $D \backslash D_1$ will be a thin annulus where we modify (1.2), (1.3) to insure that $\{X_k\}$ and $x(\cdot)$ are bounded. The actual algorithm is

$$\tilde{X}_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k \sigma(X_k) W_k$$

$$X_{k+1} = \tilde{X}_{k+1} 1_D(\tilde{X}_{k+1}) + X_k 1_{\mathbb{R}^r \backslash D}(\tilde{X}_{k+1}), \tag{2.1}$$

and the associated SDE is

$$dx(t) = -\nabla U(x(t))dt + c(t)\sigma(x(t))dw(t) . \tag{2.2}$$

We will make assumptions on $U(\cdot)$ and $\sigma(\cdot)$ to force $\{\tilde{X}_k\}$ and $x(\cdot)$ to eventually stay in D when they start in D.

In the sequel we make the following assumptions:

(A1) $U(\cdot)$ is a twice continuously differentiable function from D to $[0,\infty)$ with $\min_{x \in D} U(x) = 0$ and $\langle \nabla U(x),x \rangle > 0$ for all $x \in D \backslash \overset{\circ}{D}_1$.

(A2) $\sigma(\cdot)$ is a Lipshitz continuous function from D to [0,1] with $\sigma(x) > 0$ for all $x \in \overset{\circ}{D}$, $\sigma(x) = 1$ for all $x \in D_1$, and $\sigma(x) = 0$ for all $x \in \partial D$.

(A3) $\{\xi_k\}$ is a sequence of $\mathbb{R}^r$-valued random variables; $\{W_k\}$ is a sequence of independent r-dimensional Gaussian random variables with zero-mean and identity covariance matrix.

(A4) $a_k = \dfrac{A}{k}$, $b_k^2 = \dfrac{B}{k \log \log k}$, k large, where $A, B > 0$.

(A5) $c^2(t) = \dfrac{C}{\log t}$, t large, where $C > 0$.

For every k=0,1,... let $\mathscr{F}_k$ be the σ-field generated by $\{X_0,\xi_0,...,\xi_{k-1},W_0,...W_{k-1}\}$.

(A6) $E\{|\xi_k|^2|\mathscr{F}_k\} = O(a_k^\alpha)$, $E\{\xi_k|\mathscr{F}_k\} = O(a_k^\beta)$ and $|\xi_k|1_{D\backslash \mathring{D}_1}(X_k) \longrightarrow 0$ as $k\longrightarrow\infty$ uniformly w.p.1; W_k is independent of $\mathscr{F}_k$ for all k.

For every $\epsilon > 0$ let

$$\pi^\epsilon(x) = \frac{1}{Z^\epsilon}\exp\left(-\frac{2\,U(x)}{\epsilon^2}\right)1_D(x);\;\; Z^\epsilon = \int_D \exp\left(-\frac{2\,U(x)}{\epsilon^2}\right)dx$$

(A7) π^ϵ has a unique weak limit π as $\epsilon\longrightarrow 0$.

A few remarks about these assumptions are in order. First it is clear that π concentrates on S^*, the global minima of $U(\cdot)$. The existence of π and a simple characterization in terms of the Hessian of $U(\cdot)$ is discussed in [14]. Also, it is clear that the $P\bigcap_{t\geq 0}\{x(t)\in D\} = 1$ when $x(0)\in D$ and it can be shown that $P\bigcup_n\bigcap_{k\geq n}\{\tilde{X}_k\in D\} = 1$ when $X_o\in D$ and $\alpha > -1$ (see the Remark following Proposition 1 in Section 3). Finally, we point out that a penalty function can be added to $U(\cdot)$ so that $\nabla U(\cdot)$ points outward in the annulus $D\backslash\mathring{D}_1$ as in (A1). However, the condition that ξ_k tends to zero in the annulus $D\backslash\mathring{D}_1$ as in (A6) can be a significant restriction.

For a process $u(\cdot)$ and function $f(\cdot)$, let $E_{t_1,u_1}\{f(u(t))\}$ denote conditional expectation with respect to $u(t_1) = u_1$ and let $E_{t_1,u_1;t_2,u_2}\{f(u(t))\}$ denote conditional expectation with respect to $u(t_1) = u_1$ and $u(t_2) = u_2$. Also for a measure $\mu(\cdot)$ and a function $f(\cdot)$ let $\mu(f) = \int f d\mu$.

By a modification of the main result of [11] we have that there exists a constant C_0 such that for $C > C_0$ and any bounded and continuous function $f(\cdot)$ on $\mathbb{R}^r$

$$\lim_{t\longrightarrow\infty} E_{0,x}\{f(x(t))\} = \pi(f) \tag{2.3}$$

uniformly for $x\in D$. In [11] the constant C_0 is denoted by c_0 and has an interpretation in terms of the action functional for the dynamical system $\dot{z}(t) = -\nabla U(z(t))$. Here is our theorem on the convergence of $\{X_k\}$.

Theorem: Let $\alpha > -1$, $\beta > 0$, and $B/A > C_0$. Then for any bounded continuous function $f(\cdot)$ on $\mathbb{R}^r$

$$\lim_{k \to \infty} E_{0,x}\{f(X_k)\} = \pi(f) \tag{2.4}$$

uniformly for $x \in D$.

Since π concentrates on S^*, (2.3) and (2.4) imply $x(t) \to S^*$ and $X_k \to S^*$ in probability, respectively.

The proof of the theorem requires the following three Lemmas. Let $\{t_k\}$ and $\beta(\cdot)$ be defined by

$$t_k = \sum_{n=0}^{k-1} a_n \ , \ \ k = 0,1,...,$$

$$\int_s^{\beta(s)} \frac{\log s}{\log u} \, du = s^{2/3} \ , \ \ s > 1 \ .$$

Lemma 1: Let $\alpha > -1$, $\beta > 0$, and $B/A = C$. Then there exists $\gamma > 1$ such that for any bounded continuous function $f(\cdot)$ on $\mathbb{R}^r$

$$\lim_{n \to \infty} \sup_{k \,:\, t_n \le t_k \le \gamma t_n} E_{0,x;n,y} \ \{f(X_k)\} - E_{t_n,y}\{f(x(t_k))\} = 0$$

uniformly for $x,y \in D$.

Lemma 2: For any bounded continuous function $f(\cdot)$ on $\mathbb{R}^r$

$$\lim_{n \to \infty} \sup_{s \,:\, t_n \le s \le t_{n+1}} E_{t_n,y}\{f(x(\beta(s)))\} - E_{s,y}\{f(x(\beta(s)))\} = 0$$

uniformly for $y \in D$.

Lemma 3: Let $C > C_0$. Then for any bounded continuous function $f(\cdot)$ on $\mathbb{R}^r$

$$\lim_{s \to \infty} E_{s,y}\{f(x(\beta(s)))\} - \pi^{c(s)}(f) = 0$$

uniformly for $y \in D$.

The proofs of Lemmas 1 and 2 are in Section 3. Lemma 3 is a modification of results in [11, Lemmas 2, 3]. Note how the Lemmas are concerned with nonuniform approximation on intervals of increasing length, as opposed to uniform approximation on intervals of fixed length.

We now show how the Lemmas may be combined to prove the Theorem.

Proof of Theorem: Note that $\beta(s)$ is a strictly increasing function and $s + s^{2/3} \le \beta(s) \le s + 2s^{2/3}$ for s large enough. Hence for k large enough one can choose s such that $t_k = \beta(s)$. Clearly $s < t_k$ and $s \to \infty$ as $k \to \infty$. Furthermore for k and hence s large enough one can choose n such that $t_n \le t_k \le \gamma t_n$ and $t_n \le s \le t_{n+1}$. Clearly $n < k$ and $n \to \infty$ as $k \to \infty$. Let $p(0,x;n,A) = P\{X_n \in A | X_0 = x\}$. We can write

$$E_{0,x}\{f(X_k)\} - \pi(f) = \int_D p(0,x;n,dy)\,(E_{0,x;n,y}\{f(X_k)\} - \pi(f)). \tag{2.5}$$

Now

$$\begin{aligned} E_{0,x;n,y}\{f(X_k)\} - \pi(f) &= E_{0,x;n,y}\{f(X_k))\} - E_{t_n,y}\{f(x(t_k))\} \\ &+ E_{t_n,y}\{f(x(\beta(s)))\} - E_{s,y}\{f(x(\beta(s)))\} \\ &+ E_{s,y}\{f(x(\beta(s)))\} - \pi^{c(s)}(f) \\ &+ \pi^{c(s)}(f) - \pi(f) \to 0 \quad \text{as} \quad k \to \infty \end{aligned} \tag{2.6}$$

uniformly for $x,y \in D$ by Lemmas 1-3 and (A7). Combining (2.5) and (2.6) completes the proof.

□

As an illustration of our Theorem, we examine the random directions version of (1.2) that was implemented in [13]. If we could make noiseless measurements of $\nabla U(X_k)$ then we could use the algorithm

$$X_{k+1} = X_k - a_k \nabla U(X_k) + b_k W_k \tag{2.7}$$

(modified as in (2.1)). Suppose that $\nabla U(X_k)$ is not available but we can make noiseless measurements of $U(\cdot)$. Suppose we replace $\nabla U(X_k)$ in (2.7) by a forward finite difference approximation of $\nabla U(X_k)$, which would require $r+1$ evaluations of $U(\cdot)$. It can be shown that such an algorithm can be written in the form of (1.2) with $\xi_k = O(c_k)$ where $\{c_k\}$ are the finite difference intervals ($c_k \to 0$). As an alternative, suppose that at each iteration a direction d_k is chosen at random and we replace $\nabla U(X_k)$ in (2.7) by a finite difference approximation of the directional derivative $\langle \nabla U(X_k), d_k \rangle d_k$ in the direction d_k, which only requires 2 evaluations of $U(\cdot)$. Conceivably, fewer evaluations of $U(\cdot)$ would be required by such a random directions algorithm to converge. Now assume that the $\{d_k\}$ are random vectors each distributed uniformly over the surface of the $r-1$ dimensional sphere and that d_k is independent of $X_0, W_0, \ldots W_{k-1}, d_0, \ldots, d_{k-1}$. By analysis similar to [4, p. 58-60] it can be shown that such a random directions algorithm can be written in the form of (1.2) with $E\{\xi_k \mid \mathscr{F}_k\} = O(c_k)$ and $\xi_k = O(1)$. Hence the conditions of the Theorem will be satisfied and convergence will be obtained provided that the finite difference approximation of $\nabla U(X_k)$ is used in the thin annulus $D \backslash D_1$ and $c_k = O(k^{-\beta})$ for some $\beta > 0$.

Our Theorem, like Kushner's [12], requires that the trajectories of $\{X_k\}$ be bounded. However, there is a version of Lemma 3 in [11] which applies with $D = \mathbb{R}^r$ assuming certain growth conditions on $U(\cdot)$. We are currently trying to obtain versions of Lemmas 1 and 2 which also hold for $D = \mathbb{R}^r$. On the other hand, we have found that bounding the trajectories of $\{X_k\}$ seems useful and even necessary in practice. The reason is that even with the specified growth conditions $|X_k|$ tends occasionally to very large values which leads to numerical problems in the simulation.

There are many hard multivariate optimization problems where the simulated annealing type algorithms discussed in this paper might be applied. Recently there has been alot of interest in learning algorithms for artificial neural networks. In particular the so-called backpropagation algorithm has emerged as a popular method for training multilayer perceptron networks [15]. Backpropagation is a stochastic descent algorithm and as such is subject to getting trapped in local minima. It would be interesting to determine whether a simulated annealing type backpropagation algorithm where slowly decreasing noise has been added in artificially can alleviate this problem.

3. PROOFS OF LEMMAS 1 and 2

Throughout this section it will be convenient to make the following assumption in place of (A5):

(A5′) $c^2(t_k) = \dfrac{C}{k \log\log k}$, k large, where $C > 0$, and $c^2(\cdot)$ is a piecewise linear interpolation of $\{c^2(t_k)\}$

Note that under (A5′) $c^2(t) \sim C/\log t$ as $t \to \infty$, and if $B/A = C$ then $b_k = \sqrt{a_k}\, c(t_k)$ for k large enough. The results are unchanged whether we assume (A5) or (A5′). We shall also assume that a_k, b_k and $c(t)$ are all bounded above by 1. In the sequel $c_1, c_2, \ldots$, will denote positive constants whose value may change from proof to proof.

We start with several Propositions.

Proposition 1:

$$P\{\tilde{X}_{k+1} \notin D \,|\mathscr{F}_k\} = O(a_k^{2+\alpha}) \quad \text{as} \quad k \to \infty,$$

uniformly w.p.1.

Proof : Let $r_k = \sqrt{k}$, $k = 0, 1, \ldots$. We can then write

$$\begin{aligned} P\{\tilde{X}_{k+1} \notin D \,|\mathscr{F}_k\} &= P\{\tilde{X}_{k+1} \notin D,\ |W_k| \geq r_k \,|\mathscr{F}_k\} \\ &\quad + P\{\tilde{X}_{k+1} \notin D,\ |W_k| \leq r_k \,|\mathscr{F}_k\} 1_{\mathring{D}_1}(X_k) \\ &\quad + P\{\tilde{X}_{k+1} \notin D,\ |W_k| \leq r_k \,|\mathscr{F}_k\} 1_{D \setminus \mathring{D}_1}(X_k) \end{aligned} \qquad (3.1)$$

We bound each term on the r.h.s. of (3.1) as follows.

First, we have

$$\begin{aligned} &P\{\tilde{X}_{k+1} \notin D,\ |W_k| \geq r_k \,|\mathscr{F}_k\} \\ &\leq P\{|W_k| \geq r_k\} \leq r \exp\left(-\frac{r_k^2}{2r}\right) = o(a_k^{2+\alpha}) \quad \text{as } k \to \infty. \end{aligned} \qquad (3.2)$$

Here we have adapted the standard estimate $\Pr\{\eta > x\} \leq \frac{1}{2}\exp(-x^2/2)$ for $x \geq 0$, where η is a scalar zero-mean unit variance Gaussian random variable.

Next, we show that

$$P\{\tilde{X}_{k+1} \notin D,\ |W_k| \le r_k \,|\mathscr{F}_k\} 1_{\mathring{D}_1}(X_k) = O(a_k^{2+\alpha}) \ \text{ as } k \to \infty. \tag{3.3}$$

Let $X_k \in \mathring{D}_1$. Let $\epsilon_1 = \inf_{x \in D_1, y \in \partial D} |x-y| > 0$ and $0 < \epsilon_2 < \epsilon_1$. Then

$$P\{\tilde{X}_{k+1} \notin D,\ |W_k| \le r_k \,|\mathscr{F}_k\}$$

$$\le P\{\,|-a_k(\nabla U(X_k) + \xi_k) + b_k W_k| > \epsilon_1,\ |W_k| \le r_k \,|\mathscr{F}_k\}$$

$$\le P\{a_k |\xi_k| > \epsilon_2 \,|\mathscr{F}_k\} \le \frac{a_k^2 E\{\,|\xi_k|^2 \,|\mathscr{F}_k\}}{\epsilon_2^2} = O(a_k^{2+\alpha}) \ \text{ as } k \to \infty\,.$$

The second inequality follows from the fact that $b_k r_k \to 0$ as $k \to \infty$, and the third inequality is Chebyshev's. This proves (3.3).

Finally, we show that

$$P\{\tilde{X}_{k+1} \notin D,\ |W_k| \le r_k \,|\mathscr{F}_k\} 1_{D \backslash \mathring{D}_1}(X_k) = 0 \tag{3.4}$$

for k large enough. Let $X_k \in D \backslash \mathring{D}_1$. Let $\overline{X}_k = X_k + b_k \sigma(X_k) W_k 1_{\{|W_k| \le r_k\}}$. Since $\sigma(\cdot)$ is Lipshitz, $\sigma(x) > 0$ for all $x \in \mathring{D}$, and $\sigma(x) = 0$ for all $x \in \partial D$, we have $\sigma(x) \le c_1 \inf_{y \in \partial D} |x-y|$ for all $x \in D$. Hence $|\overline{X}_k - X_k| \le b_k r_k c_1 \inf_{y \in \partial D} |X_k - y|$, and since $b_k r_k \to 0$ as $k \to \infty$ we get $\overline{X}_k - X_k \to 0$ as $k \to \infty$ and $\overline{X}_k \in D$ for k large enough. Now since $X_k \in D \backslash \mathring{D}_1$ we have $<\nabla U(X_k), X_k> \,> c_2$ and $\xi_k \to 0$ as $k \to \infty$. Hence $<\nabla U(X_k) + \xi_k, \overline{X}_k> - <\nabla U(X_k) + \xi_k, X_k> \to 0$ as $k \to \infty$ and $<\nabla U(X_k) + \xi_k, X_k> \,> c_2 > 0$ for k large enough, and so $<\nabla U(X_k) + \xi_k, \overline{X}_k> \,> c_2 > 0$ for k large enough, and consequently

$$\frac{<a_k(\nabla U(X_k) + \xi_k), \overline{X}_k>}{|a_k(\nabla U(X_k) + \xi_k)|\,|\overline{X}_k|} > c_3 > 0$$

for k large enough. But $\tilde{X}_{k+1} = \overline{X}_k - a_k(\nabla U(X_k) + \xi_k) \in D$ whenever $\overline{X}_k \in D$ and $|a_k(\nabla U(X_k) + \xi_k)| \le c_3 \cdot \operatorname{diam} D$, and these hold for k large enough. This proves (3.4). Combining (3.1)-(3.4) to completes the proof.

□

Remark: By Proposition 1 and the Borel-Cantelli Lemma $P \bigcup_n \bigcap_{k \geq n} \{\tilde{X}_k \in D\} = 1$ when $X_0 \in D$ and $\alpha > -1$.

Proposition 2: For each n let $\{u_{n,k}\}_{k \geq n}$ be a sequence of nonnegative numbers such that

$$u_{n,k+1} \leq (1 + ca_k)u_{n,k} + ca_k^\delta, \qquad k \geq n,$$

$$u_{n,n} = O(a_n^\epsilon) \text{ as } n \rightarrow \infty,$$

where $\delta > 1$, $\epsilon > 0$, and $c > 0$. Then there exists a $\gamma > 1$ such that

$$\lim_{n \rightarrow \infty} \sup_{k:t_n \leq t_k \leq \gamma t_n} u_{n,k} = 0.$$

Proof : We may set c=1 since $a_k = A/k$ for k large and $A > 0$ is arbitrary. Now

$$u_{n,k} \leq u_{n,n} \prod_{\ell=n}^{k-1} (1+a_\ell) + \sum_{m=n}^{k-1} a_m^\delta \prod_{\ell=m+1}^{k-1} (1+a_\ell)$$

$$\leq (u_{n,n} + \sum_{m=n}^{k-1} a_m^\delta) \cdot \exp(\sum_{m=n}^{k-1} a_m),$$

since $1 + x \leq e^x$ for all x. Also $\Sigma_n^{k-1} a_m \leq A(\log(k/n)+1/n)$ and $\Sigma_n^{k-1} a_m^\delta \leq A(1/(\delta-1)n^{\delta-1} + 1/n^\delta)$, and if $t_k \leq \gamma t_n$ then $k \leq c_1 n^\gamma$. Choose γ such that $1 < \gamma < 1 + \min\{\delta-1, \epsilon\}/A$. It follows that

$$\sup_{k:t_n \leq t_k \leq \gamma t_n} u_{n,k} \leq c_2 \left(\frac{1}{n^\epsilon} + \frac{1}{n^{\delta-1}} \right) n^{(\gamma-1)A} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

□

Define $\xi(\cdot,\cdot)$ by

$$x(t) = x(s) - (t-s)(\nabla U(x(s)) + \xi(s,t)) + c(s)\sigma(x(s))(w(t) - w(s))$$

for $t \geq s \geq 0$.

Proposition 3:

$$E\{\xi(t,t+h) \,|x(t)\} = O(h^{1/2}),$$

$$E\{\,|\xi(t,t+h)\,|^2 \,|x(t)\} = O(1),$$

as $h \rightarrow 0$, uniformly for a.e. $x(t) \in D$ and all $t \geq 0$.

Proof : We use some elementary facts about stochastic integrals and martingales (c.f. [16]). First write

$$h\xi(t,t+h) = \int_t^{t+h} (\nabla U(x(\tau)) - \nabla U(x(t))d\tau$$

$$- \int_t^{t+h} (c(\tau)\sigma(x(\tau)) - c(t)\sigma(x(t)))dw(\tau) \tag{3.5}$$

Now a standard result is that

$$E\{ |x(t+h) - x(t)|^2 \, |x(t)\} = O(h)$$

as $h \longrightarrow 0$, uniformly for a.e. $x(t) \in D$ and t in a finite interval. In fact, under our assumptions the estimate is uniform here for a.e. $x(t) \in D$ and all $t \geq 0$. Let K_1,K_2 be Lipshitz constants for $\nabla U(\bullet)$, $\sigma(\bullet)$, respectively. Also note that $c(\bullet)$ is piecewise continuously differentiable with bounded derivative (where it exists) and hence is also Lipshitz continuous, say with constant K_3. Hence

$$E\{ |\int_t^{t+h} (\nabla U(x(\tau)) - \nabla U(x(t)))d\tau|^2 \, |x(t)\}$$

$$\leq K_1^2 E\{(\int_t^{t+h} |x(\tau) - x(t)|d\tau)^2 \, |x(t)\}$$

$$\leq K_1^2 h \int_t^{t+h} E\{ |x(\tau) - x(t)|^2 \, |x(t)\}d\tau = O(h^3) \tag{3.6}$$

and

$$E\{ |\int_t^{t+h} (c(\tau)\sigma(x(\tau)) - c(t)\sigma(x(t)))dw(\tau)|^2 \, |x(t)\}$$

$$= \int_t^{t+h} E\{ |c(\tau)\sigma(x(\tau)) - c(t)\sigma(x(t))|^2 \, |x(t)\}d\tau$$

$$\leq 2K_2^2 \int_t^{t+h} E\{ |x(\tau) - x(t)|^2 \, |x(t)\}d\tau + 2K_3^2 \int_t^{t+h} (\tau-t)^2 d\tau = O(h^2) \tag{3.7}$$

as $h \longrightarrow 0$, uniformly for a.e. $x(t) \in D$ and all $t \geq 0$. The Proposition follows easily from

(3.5)-(3.7) and the fact that the second (stochastic) integral in (3.5) defines a martingale as h varies.

□

Now in Lemma 1 we compare the distributions of X_k and $x(t_k)$. This is done most easily by comparing X_k and $x(t_k)$ to Y_k and $\overline{Y}_k$ (defined below), respectively, which are equal in distribution.

Let

$$\tilde{Y}_{k+1} = Y_k - a_k \nabla U(Y_k) + b_k \sigma(Y_k) W_k$$

$$Y_{k+1} = \tilde{Y}_{k+1} 1_D(\tilde{Y}_{k+1}) + Y_k 1_{\mathbb{R}^r \setminus D}(\tilde{Y}_{k+1})$$

Lemma 1.1: There exists $\gamma > 1$ such that for any bounded and continuous function $f(\cdot)$ on $\mathbb{R}^r$

$$\lim_{n\to\infty} \sup_{k:t_n \le t_k \le \gamma t_n} E_{0,x;n,y}\{f(X_k)\} - E_{n,y}\{f(Y_k)\} = 0,$$

uniformly for $x,y \in D$

Proof: Let $x,y \in D$, n a positive integer, $X_0 = x$, and $X_n = Y_n = y$. Let $\Delta_k = X_k - Y_k$ for $k \ge n$. We suppress the dependence of Δ_k on x, y and n. Write

$$E\{|\Delta_{k+1}|^2\} = E\{|\Delta_{k+1}|^2 1_{\{\tilde{X}_{k+1} \notin D\} \cup \{\tilde{Y}_{k+1} \notin D\}}\}$$

$$+ E\{|\Delta_{k+1}|^2 1_{\{\tilde{X}_{k+1} \in D\} \cap \{\tilde{Y}_{k+1} \in D\}}\} \tag{3.8}$$

We estimate the first term in (3.8) as follows. We have by Proposition 1 that

$$E\{|\Delta_{k+1}|^2 1_{\{\tilde{X}_{k+1} \notin D\} \cup \{\tilde{Y}_{k+1} \notin D\}}\}$$

$$\le c_1(P\{\tilde{X}_{k+1} \notin D\} + P\{\tilde{Y}_{k+1} \notin D\}) = O(a_k^{2+\alpha}) \text{ as } k \to \infty, \tag{3.9}$$

uniformly for $x,y \in D$.

We estimate the second term in (3.8) as follows. If $\tilde{X}_{k+1} \in D$ and $\tilde{Y}_{k+1} \in D$ then

$$\Delta_{k+1} = \Delta_k - a_k(\nabla U(Y_k + \Delta_k) - \nabla U(Y_k))$$

$$+ b_k(\sigma(Y_k + \Delta_k) - \sigma(Y_k))W_k - a_k \xi_k.$$

Hence

$$
\begin{aligned}
&E\{|\Delta_{k+1}|^2 1_{\{\tilde{X}_{k+1}\in D\}\cap\{\tilde{Y}_{k+1}\in D\}}\} \\
&\quad \le E\{|\Delta_k - a_k(\nabla U(Y_k+\Delta_k) - \nabla U(Y_k)) \\
&\qquad + b_k(\sigma(Y_k+\Delta_k) - \sigma(Y_k))W_k - a_k\xi_k|^2\} \\
&\quad \le E\{|\Delta_k|^2\} + a_k^2 E\{|\nabla U(Y_k+\Delta_k) - \nabla U(Y_k)|^2\} \\
&\qquad + a_k E\{|(\sigma(Y_k+\Delta_k) - \sigma(Y_k))W_k|^2\} \\
&\qquad + a_k^2 E\{|\xi_k|^2\} \\
&\qquad + 2a_k |E\{<\Delta_k, \nabla U(Y_k+\Delta_k) - \nabla U(Y_k)>\}| \\
&\qquad + 2a_k^{1/2} |E\{<\Delta_k, (\sigma(Y_k+\Delta_k) - \sigma(Y_k))W_k>\}| \\
&\qquad + 2a_k |E\{<\Delta_k, \xi_k>\}| \\
&\qquad + 2a_k^{3/2} |E\{<\nabla U(Y_k+\Delta_k) - \nabla U(Y_k), (\sigma(Y_k+\Delta_k) - \sigma(Y_k))W_k>\}| \\
&\qquad + 2a_k^2 |E\{<\nabla U(Y_k+\Delta_k) - \nabla U(Y_k),\ \xi_k>\}| \\
&\qquad + 2a_k^{3/2} |E\{<(\sigma(Y_k+\Delta_k) - \sigma(Y_k))W_k, \xi_k>\}|, \qquad (3.10)
\end{aligned}
$$

for all $x,y\in D$, $k \ge n$, and n large enough. Let K_1, K_2 be Lipshitz constants for $\nabla U(\cdot)$, $\sigma(\cdot)$, respectively. Using the facts that X_k, Y_k and hence Δ_k are $\mathscr{F}_k$ measurable, W_k is independent of $\mathscr{F}_k$, and

$$|E\{|\xi_k|^2 | \mathscr{F}_k\} \le c_2 a_k^{\alpha}, \qquad |E\{\xi_k | \mathscr{F}_k\}| \le c_2 a_k^{\beta},$$

w.p.1 for all $x,y\in D$, $k\ge n$, and n large enough, we have

$$E\{|\nabla U(Y_k+\Delta_k) - \nabla U(Y_k)|^2\} \le K_1^2 E\{|\Delta_k|^2\}$$

$$E\{|(\sigma(Y_k+\Delta_k) - \sigma(Y_k))W_k|^2\} \le rK_2^2 E\{|\Delta_k|^2\}$$

$$E\{|\xi_k|^2\} \le c_2 a_k^{\alpha}$$

$$|E\{<\Delta_k, \nabla U(Y_k+\Delta_k) - \nabla U(Y_k)>\}| \le K_1 E\{|\Delta_k|^2\}$$

$$|E\{<\Delta_k,(\sigma(Y_k+\Delta_k) - \sigma(Y_k))W_k\}|$$

$$= |E\{<\Delta_k,(\sigma(Y_k+\Delta_k) - \sigma(Y_k))E\{W_k\}>\}| = 0$$

$$|E\{<\Delta_k,\xi_k>\}| = |E\{<\Delta_k,E\{\xi_k|\mathscr{F}_k\}>|$$

$$\leq E\{|\Delta_k||E\{\xi_k|\mathscr{F}_k\}|\} \leq c_2 a_k^\beta E\{|\Delta_k|\}$$

$$|E\{<\nabla U(Y_k+\Delta_k) - \nabla U(Y_k),\ (\sigma(Y_k+\Delta_k) - \sigma(Y_k))W_k>\}|$$

$$\leq |E\{<\nabla U(Y_k+\Delta_k) - \nabla U(Y_k),(\sigma(Y_k+\Delta_k) - \sigma(Y_k))E\{W_k\}>\}| = 0$$

$$|E\{<\nabla U(Y_k+\Delta_k)-\nabla U(Y_k),\xi_k>\}| = |E\{<\nabla U(Y_k+\Delta_k)-\nabla U(Y_k),E\{\xi_k|\mathscr{F}_k\}>\}|$$

$$= E\{|\nabla U(Y_k+\Delta_k) - \nabla U(Y_k)||E\{\xi_k|\mathscr{F}_k\}|\} \leq c_2 K_1 a_k^\beta E\{|\Delta_k|\}$$

$$|E\{<(\sigma(Y_k+\Delta_k) - \sigma(Y_k))W_k,\xi_k>\}| = |E\{(\sigma(Y_k+\Delta_k) - \sigma(Y_k))E\{<W_k,\xi_k>|\mathscr{F}_k\}|$$

$$\leq E\{|\sigma(Y_k+\Delta_k)-\sigma(Y_k)|E\{|W_k|^2\}^{1/2}E\{|\xi_k|^2|\mathscr{F}_k\}^{1/2}|\} \leq \sqrt{rc_2}K_2 a_k^{\alpha/2}E\{|\Delta_k|\}$$

for all $x,y\in D$, $k\geq n$, and n large enough. Substituting these expressions into (3.10) gives (after some simplification)

$$\begin{aligned} E\{|\Delta_{k+1}|^2 1_{\{\tilde{X}_{k+1}\in D\}\cap\{\tilde{Y}_{k+1}\in D\}}\} &\leq (1+c_3 a_k)E\{|\Delta_k|^2\} + c_3 a_k^{\delta_1}E\{|\Delta_k|\} + c_2 a_k^{2+\alpha} \\ &\leq (1+c_3 a_k)E\{|\Delta_k|^2\} + c_3 a_k^{\delta_1}E\{|\Delta_k|^2\}^{1/2} + c_2 a_k^{2+\alpha} \\ &\leq (1+c_4 a_k)E\{|\Delta_k|^2\} + c_4 a_k^{\delta_2}, \end{aligned} \tag{3.11}$$

for all $x,y\in D$, $k\geq n$, and n large enough, where $\delta_1 = \min\{1+\beta,(3+\alpha)/2\}>1$ and $\delta_2 = \min\{\delta_1,2+\alpha\}>1$ since $\alpha>-1$ and $\beta>0$.

Now combine (3.8), (3.9) and (3.11) to get

$$E\{|\Delta_{k+1}|^2\} \leq (1 + c_5 a_k)E\{|\Delta_k|^2\} + c_5 a_k^{\delta_2},\quad k\geq n,$$

$$E\{|\Delta_n|^2\} = 0,$$

for all $x,y\in D$ and n large enough. Applying Proposition 2 there exists $\gamma>1$ such that

$$\lim_{n\to\infty} \sup_{k:t_n\le t_k\le \gamma t_n} E\{\,|\Delta_k|^2\} = 0, \tag{3.12}$$

uniformly for all $x,y\in D$.

Finally, let $f(\cdot)$ be a bounded continuous function on $\mathbb{R}^r$. Since D is compact $f(\cdot)$ is uniformly continuous on D. So given $\epsilon>0$ let $\delta>0$ be such that $|f(u)-f(v)|<\epsilon$ whenever $|u-v|<\delta$ and $u,v\in D$. Then

$$\begin{aligned}|E_{0,x;n,y}\{f(X_k)\} - E_{n,y}\{f(Y_k)\}\,| &\le \epsilon P\{\,|\Delta_k| < \delta\} + 2||f||P\{\,|\Delta_k|>\delta\} \\ &\le \epsilon + \frac{2||f||}{\delta^2}E\{\,|\Delta_k|^2\},\end{aligned}$$

and by (3.12)

$$\overline{\lim_{n\to\infty}} \sup_{k:t_n\le t_k\le \gamma t_n} |E_{0,x;n,y}\{f(X_k)\} - E_{n,y}\{f(Y_k)\}\,| \le \epsilon,$$

uniformly for $x,y\in D$, and letting $\epsilon\to 0$ completes the proof.

□

Let $\overline{W}_k = (w(t_{k+1})-w(t_k))/\sqrt{a_k}$ and

$$\tilde{\overline{Y}}_{k+1} = \overline{Y}_k - a_k\nabla U(\overline{Y}_k) + b_k\sigma(\overline{Y}_k)\overline{W}_k$$

$$\overline{Y}_{k+1} = \tilde{\overline{Y}}_{k+1}1_D(\tilde{\overline{Y}}_{k+1}) + \overline{Y}_k 1_{\mathbb{R}^r\setminus D}(\tilde{\overline{Y}}_{k+1})$$

Lemma 1.2: There exists $\gamma>1$ such that for any bounded continuous function $f(\cdot)$ on $\mathbb{R}^r$

$$\lim_{n\to\infty} \sup_{k:t_n\le t_k\le \gamma t_n} E_{n,y}\{f(x(t_k))\} - E_{n,y}\{f(\overline{Y}_k)\} = 0$$

uniformly for $y\in D$.

Proof : Let $y\in D$, n be a positive integer, and $x(t_n) = \overline{Y}_n = y$. Define $\{\overline{\xi}_k\}$ by

$$x(t_{k+1}) = x(t_k) - a_k(\nabla U(x(t_k)) + \overline{\xi}_k) + b_k\sigma(x(t_k))\overline{W}_k, \qquad k\ge n.$$

Let $\overline{\mathscr{F}}_k$ be the σ-field generated by $\{x(t_n),\overline{\xi}_n,\dots,\overline{\xi}_{k-1},\overline{W}_n,\dots,\overline{W}_{k-1}\}$ for $k\ge n$. It can be shown that $\overline{\xi}_k$ is conditionally independent of $\overline{\mathscr{F}}_k$ given $x(t_k)$. Hence by Proposition 3

$$E\{|\bar\xi_k|^2|\mathscr{F}_k\} \le c_1, \qquad |E\{\bar\xi_k|\mathscr{F}_k\}| \le c_1 a_k^{1/2},$$

w.p.1 for all $y \in D$, $k \ge n$, and n large enough. Let $\Delta_k = x(t_k) - \bar Y_k$ for $k \ge n$. We suppress the dependence of Δ_k on y and n. Similiarly to the proof of Lemma 1.1 we can show with $\delta = 3/2$ that

$$E\{|\Delta_{k+1}|^2\} \le (1+c_2 a_k)E\{|\Delta_k|^2\} + c_2 a_k^{\delta}, \quad k \ge n,$$

$$E\{|\Delta_n|^2\} = 0,$$

for all $y \in D$ and n large enough. Applying Proposition 2 there exists a $\gamma > 1$ such that

$$\lim_{n\to\infty} \sup_{k:t_n \le k \le \gamma t_n} E\{|\Delta_k|^2\} = 0,$$

uniformly for $y \in D$. The Lemma now follows as in the proof of Lemma 1.1.

□

Proof of Lemma 1: Follow immediately from Lemmas 1.1 and 1.2.

□

Proof of Lemma 2: Let $y \in D$, n a positive integer, and $s \in [t_n, t_{n+1}]$. Let $x(\cdot;s,y)$ denote the process $x(\cdot)$ emitted from y at time s. Let $v(\cdot)$ be a standard r-dimensional Wiener process starting at time t_n and independent of $x(s;t_n,y)$. Define $x_i(\cdot)$, $i = 1,2$, by

$$dx_i(t) = -\nabla U(x_i(t))dt + c(t)\sigma(x_i(t))dv(t), \quad t \ge s,$$

$$x_1(s) = x(s;t_n,y),$$

$$x_2(s) = y.$$

Let $V_k = (v(t_{k+1}) - v(t_k))/\sqrt{a_k}$ for $k > n$, and $V_n = (v(t_{n+1}) - v(s))/\sqrt{t_{n+1} - s}$. Define $\{\xi_{i,k}\}$, $i = 1,2$, by

$$x_i(t_{k+1}) = x_i(t_k) - a_k(\nabla U(x_i(t_k)) + \xi_{i,k}) + b_k\sigma(x_i(t_k))V_k, \quad k > n,$$

$$x_i(t_{n+1}) = x_i(s) - (t_{n+1} - s)(\nabla U(x_i(s)) + \xi_{i,n}) + \sqrt{t_{n+1} - s}\, c(s)\sigma(x_i(s))V_n.$$

Let $\mathscr{F}_{i,k}$ be the σ-field generated by $\{x_i(s), \xi_{i,n}, \ldots, \xi_{i,k-1}, V_n, \ldots, V_{k-1}\}$ for $k \ge n$. It can be shown that $\xi_{i,k}$ is conditionally independent of $\mathscr{F}_{1,k} \vee \mathscr{F}_{2,k}$ given $x_i(t_k)$. Hence by Proposition 3

$$E\{|\xi_{1,k}+\xi_{2,k}|^2 \; |\mathscr{F}_{1,k}\vee \mathscr{F}_{2,k}\}| \leq c_1, \quad |E\{\xi_{1,k}+\xi_{2,k} \; |\mathscr{F}_{1,k}\vee \mathscr{F}_{2,k}\}| \leq c_1 a_k^{1/2},$$

w.p.1 for all $y\in D$, $s\in[t_n,t_{n+1}]$, $k\geq n$, and n large enough.

Now observe that

$$E\{|x(t+h)-x(t)|^2 \, |x(t)\} = O(h) \text{ as } h\to 0,$$

uniformly for a.e. $x(t)\in D$ and all $t\geq 0$ (this is a standard result expect for the uniformity for all t which was remarked on in Proposition 3). Hence

$$E\{|x_1(s)-x_2(s)|^2\} = E\{|x(s;t_n,y)-y|^2\} \leq c_2 a_n,$$

for all $y\in D$, $s\in[t_n,t_{n+1}]$, and n large enough. Let $\Delta_k = x_1(t_{k+1})-x_2(t_{k+1})$ for $k\geq n$. We suppress the dependence of Δ_k on y, s and n. Similiarly to the proof of Lemma 1.1 we can show with $\delta = 3/2$ that

$$E\{|\Delta_{k+1}|^2\} \leq (1+c_3 a_k)E\{|\Delta_k|^2\} + c_3 a_k^\delta, \qquad k\geq n,$$

$$E\{|\Delta_n|^2\} \leq (1+c_3 a_n)E\{|x_1(s)-x_2(s)|^2\} + c_3 a_n^\delta \leq c_4 a_n,$$

for all $y\in D$, $s\in[t_n,t_{n+1}]$, and n large enough. Hence

$$\sup_{s:t_n\leq s\leq t_{n+1}} E\{|\Delta_{k+1}|^2\} \leq (1+c_3 a_k) \sup_{s:t_n\leq s\leq t_{n+1}} E\{|\Delta_k|^2\} + c_3 a_k^\delta, \qquad k\geq n,$$

for all $y\in D$ and n large enough, and

$$\sup_{s:t_n\leq s\leq t_{n+1}} E\{|\Delta_n|^2\} = O(a_n) \text{ as } n\to\infty,$$

uniformly for $y\in D$. Applying Proposition 2 there exists $\gamma>1$ such that

$$\lim_{n\to\infty} \sup_{k:t_n\leq k\leq \gamma t_n} \sup_{s:t_n\leq s\leq t_{n+1}} E\{|\Delta_k|^2\} = 0, \tag{3.13}$$

uniformly for $y\in D$.

Note that $\beta(s)$ is a strictly increasing function of s and $s+s^{2/3}\leq \beta(s)\leq s+2s^{2/3}$ for s large enough. Hence for n large enough one can choose s such that $t_n\leq s\leq t_{n+1}$ and m such that $t_m\leq \beta(s)\leq t_{m+1}$ and $t_n\leq t_m\leq \gamma t_n$. As above we can show

$$E\{|x_1(\beta(s))-x_2(\beta(s))|^2\} \leq (1+c_3 a_m)E\{|\Delta_m|^2\} + c_3 a_m^\delta$$

$$\leq c_5 \sup_{k:t_n \leq t_k \leq \gamma t_n} E\{|\Delta_k|^2\} + c_3 a_n^\delta, \tag{3.14}$$

for all $y \in D$, $s \in [t_n, t_{n+1}]$, and n large enough. Combining (3.13), (3.14) gives

$$\lim_{n \to \infty} \sup_{s:t_n \leq s \leq t_{n+1}} E\{|x_1(\beta(s)) - x_2(\beta(s))|^2\} = 0,$$

uniformly for $y \in D$. Finally since $x_1(\beta(s))$, $x_2(\beta(s))$ are equal in distribution to $x(\beta(s);t_n,y)$, $x(\beta(s);s,y)$, respectively, the Lemma now follows as in the proof of Lemma 1.1.

□

6. REFERENCES

[1] Wasan, M.T., *Stochastic Approximation,* Cambridge University Press, 1969.

[2] Rubinstein, R. Y., *Simulation and the Monte Carlo Method,* Wiley, 1981.

[3] Ljung, L., "Analysis of recursive stochastic algorithms," *IEEE Transactions on Automatic Control,* Vol. AC-22, pp. 551-575, 1977.

[4] Kushner, H. and Clark, D., *Stochastic Approximation Methods for Constrained and Unconstrained Systems,* Applied Math. Science Series 26, Springer, Berlin, 1978.

[5] Metivier, M. and Priouret, P., "Applications of a Kushner and Clark lemma to general classes of stochastic algorithm," *IEEE Transactions on Information Theory,* Vol. IT-30, pp. 140-151, 1984.

[6] Grenender, U., *Tutorial in Pattern Theory,* Div. Applied Mathematics, Brown Univ., Providence, RI, 1984.

[7] Geman, S. and Hwang, C.R., "Diffusions for global optimization," *SIAM Journal Control and Optimization,* 24, pp. 1031-1043, 1986.

[8] Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P., "Optimization by Simulated Annealing," *Science,* 220, pp. 621-680, 1983.

[9] Gelfand, S.B., *Analysis of Simulated Annealing Type Algorithms,* Ph.D. Thesis, Dept. Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[10] Gidas, B., "Global optimization via the Langevin equation," *Proc. IEEE Conference on Decision and Control,* 1985.

[11] Chiang, T.S., Hwang, C.R. and Sheu, S.J., "Diffusion for global optimization in $\mathbb{R}^n$," *SIAM Journal Control and Optimization,* 25, pp. 737-752, 1987.

[12] Kushner, H.J., "Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via Monte Carlo," *SIAM Journal Applied Mathematics,* 47, pp. 169-185, 1987.

[13] Aluffi-Pentini, F., Parisi, V., and Zirilli, F., "Global Optimization and Stochastic Differential Equations," *Journal of Optimization Theory and Applications,* 47, pp. 1-16, 1985.

[14] Hwang, C.-R., "Laplaces method revisited: weak convergence of probability measures," *Annals of Probability,* 8, pp. 1177-1182, 1980.

[15] Rumelhart, D.E. and McClelland, J., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1,* MIT Press, Cambridge, MA, 1986.

[16] Gikhman, I.I. and Skorohod, A.V., *Stochastic Differential Equations,* Springer-Verlag, 1972.